Machine Learning en la empresa

Néstor Alonso

9 Jan 2017

Agenda

Presentación y contexto

Ciclo de vida de un modelo

Retos del ML en la empresa

Streaming Analytics

Presentación y contexto

Admiral Seguros: seguros para vehículos comercializados online y/o por teléfono.

Grupo británico, con presencia en UK, ESP, ITA, FRA, US, y algún país más. Cotiza en la bolsa de Londres, está dentro del índice FTSE100.

Un poco más en detalle:

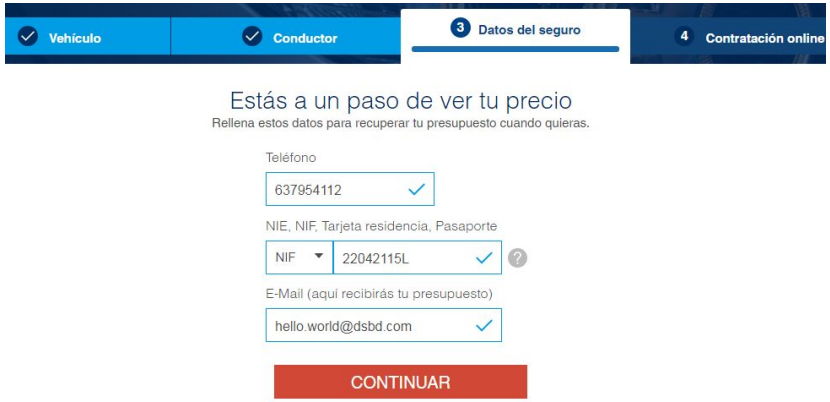
- Autos, motos y furgonetas.
- 3 marcas: Balumba, Qualitas Auto y Wiyou (no tienen que ver con los tipos de vehículos!).
- Formas de captar nuevos clientes:
  - Online (páginas web desde las que se puede cotizar, contratar, gestionar la póliza, ...).
  - Teléfono (call centre propio, personal de Admiral Seguros, 100% en Sevilla).
  - Comparadores de precios:
    - Principal canal para nosotros (Rastreator, Acierto, Kelisto, …).
    - En 5 minutos se obtienen precios de >20 cías → Alta competitividad.
    - Enorme oportunidad para ML.

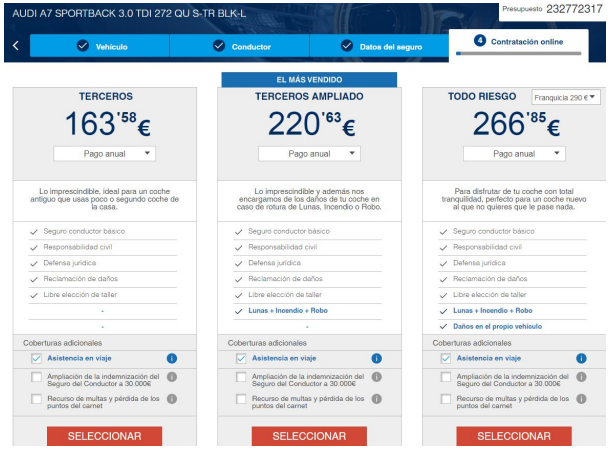
Otros grandes mundos:

- Siniestros => segmentación, fraude, optimización de procesos, eficiencia, …
- Renovación => la póliza de autos se renueva todos los años.
- Precio (la cuantificación del riesgo).

Presentación y contexto

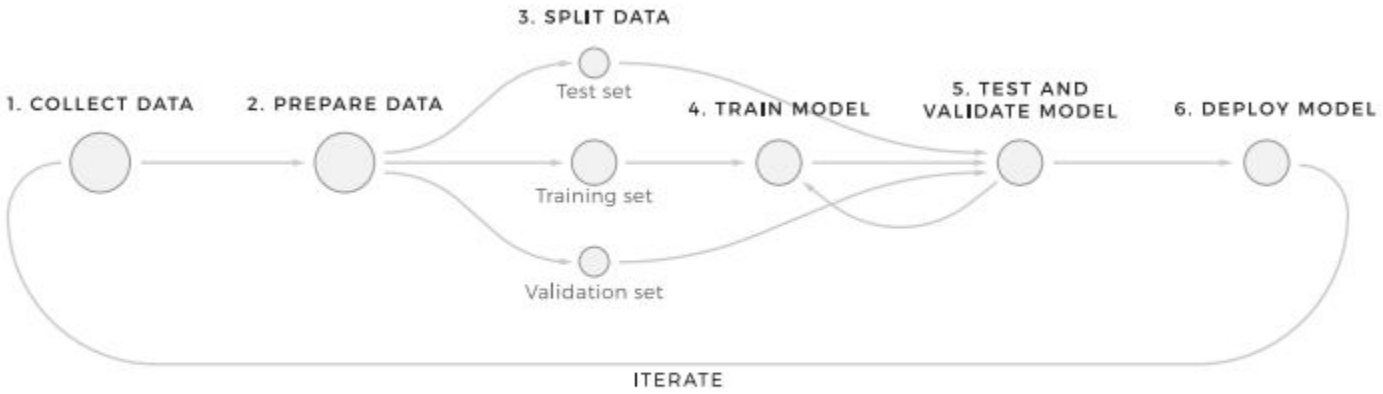
"Nuestros compromisos son la mejor prueba de nuestra eficacia".





Ciclo de vida de un modelo

El clásico diagrama:

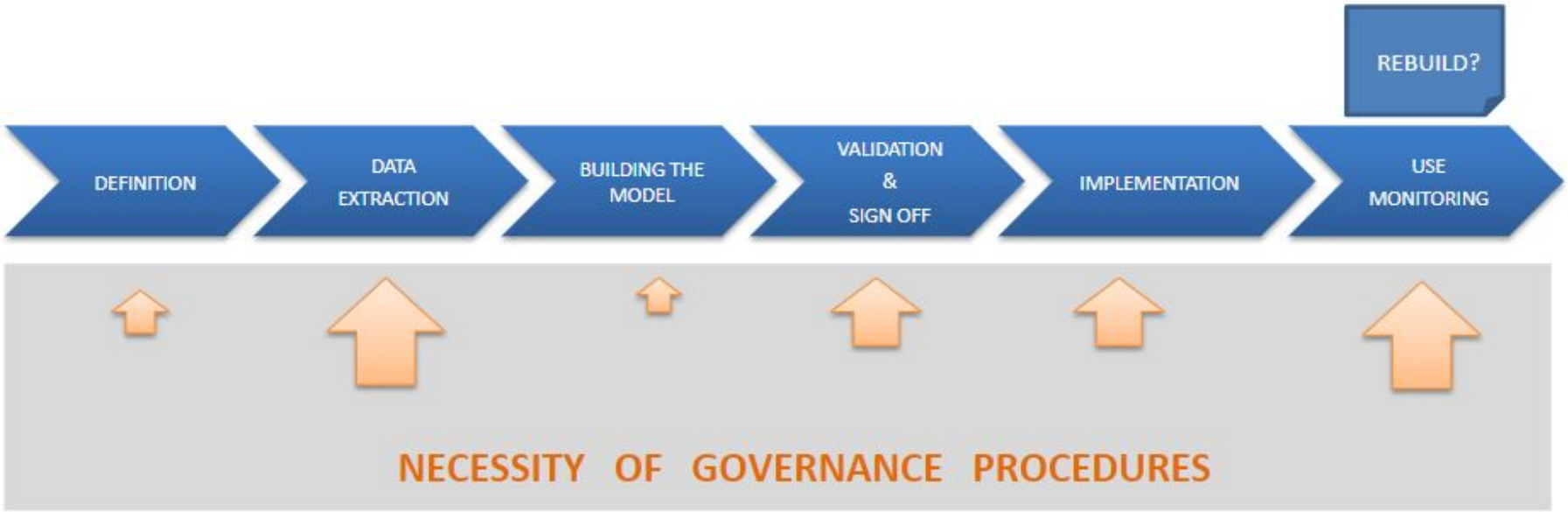


Machine Learning Workflow

Ciclo de vida de un modelo

Introducimos algunos elementos típicos en un entorno empresarial:

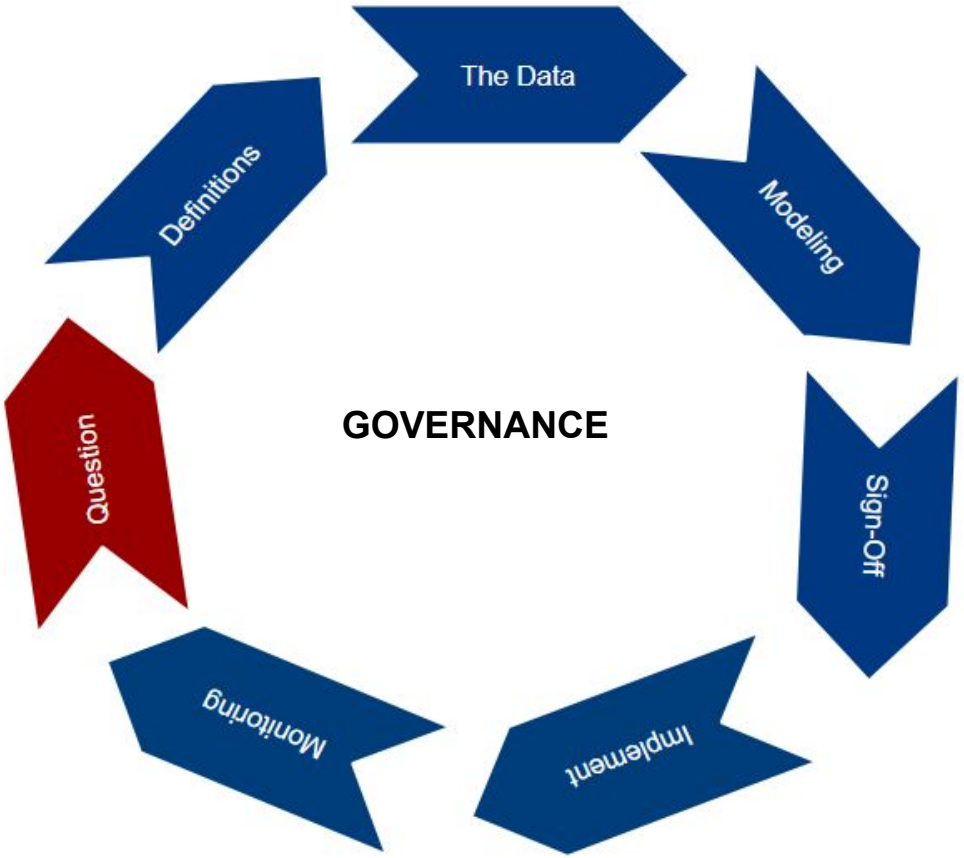
- Definición.
- Gobernanza del modelo y del ecosistema de modelos.



Ciclo de vida de un modelo

Ponemos de manifiesto el carácter cíclico.

Qué etapas dependen exclusivamente del equipo de Analytics / ML / DS, y qué etapas dependen también de otros departamentos de la empresa?



Retos del Machine Learning en la empresa

El ciclo de vida del ML supone ya de por sí un reto, y en el mundo empresarial surgen retos adicionales:

El (éxito de) ML no depende sólo del equipo de Analytics:

- ***Evangelización***… porque hay tendencias naturales:
  - Al ***agnosticismo***.
  - Al **rechazo** (resistencia al cambio, percepción de “enemigo”)

- ***Labor comercial***: tenemos que *vender* el ML y sus beneficios.

- ***Honestidad*** y la ***comunicación***:
  - Ser conscientes y comunicar los riesgos asociados al uso de ML.
  - Van a existir errores, como con cualquier otro desarrollo, herramienta, novedad...

- ***Involucrar en el proceso al resto de departamentos***:
  - Hacerles sentir parte del proceso.
  - Incorporar su conocimiento al modelo.

Quienes trabajan en empresas u organizaciones en las que el ML es el núcleo del negocio (y no un medio para lograr el éxito del negocio) posiblemente no tienen que hacer frente a los anteriores retos, así que son muy afortunados!! Pero quizá sí tengan que hacer frente a algunos de los próximos.

Admiral

Retos del Machine Learning en la empresa

Gobernanza: "El conjunto de procedimientos y herramientas que minimizan errores y facilitan el control de cada fase del ciclo, y la coexistencia *pacífica* de diferentes modelos en el *ecosistema*".

Es posiblemente el ámbito más mancillado en toda empresa, lo cual supone un gran riesgo.

Requiere muchos recursos, o bien invertir en herramientas *(existentes o propias)* para ganar eficiencia.

- ***Calidad de los datos***
  - Distintas fuentes de datos, tanto internas como externas
  - Desde los Bancos de Variables (interno)… Hacia el Data Lake (Azure, AWS, …)
  - Puede constituir un bloque en sí mismo.

- ***Legalidad de los datos*** *(podemos usar estos datos? LOPD, GDPR, …)*

- ***Usabilidad de los datos*** *(cuando esté usando este modelo, podré usar todos estos datos?)*
  - Uso de modelos en *"real-time"* vs *"batch"*

- ***Validación técnica del modelo***: *(es mi modelo generalizable a nuevos datos?)*
  - Luchar contra el sobreajuste u overfitting: C-V, hold-out, out-of-time, ...
  - Distribución del error:
    - En técnicas clásicas se puede usar Inferencia Estadística.
    - Técnicas más complejas requieren otras metodologías (bootstrapping, simulaciones).

- ***Validación funcional del modelo***: *(quién aprueba el modelo?)*
  - La aprobación debería ser de máximo nivel, involucrando al negocio.

Retos del Machine Learning en la empresa

(... seguimos con la gobernanza …)

- ***Implementación del modelo:*** *ya tengo mi modelo, ahora quiero usarlo. Tengo que ser capaz de:*
  - Acceder a los mismos datos que cuando lo construí.
  - Hacer los mismos preprocesamientos a los datos que cuando lo construí.
  - Implementar bien el algoritmo!
    - Transcribir el algoritmo:
      - Si se puede expresar como fórmula puede ser sencillo *(eg GLMs)*
      - En otros casos puede ser difícil *(eg XGBoost: xgb.dump())*
      - Especialmente difícil si el lenguaje de construcción y el de implementación no son el mismo *(eg R y Java)*.
    - Uso de máquinas virtuales (propias o externas):
      - Propias: ojo a la latencia y a la escalabilidad...
      - Externas: hacen la vida más fácil a cambio de un precio *(eg AWS, MS Azure, …)*
    - La virtualización y APIficación parece ser la tendencia: uso de algoritmos hechos con cualquier plataforma, software, lenguaje… desde otra plataforma, que accede a una versión "virtualizada" (*eg dockerizada*) del mismo.

- ***Consumo del modelo:*** *(ya tengo mi modelo en Producción! Lo estamos usando bien?)*
  - Es un proceso de negocio, pero Analytics / ML / DS debe apoyarlo y participar.
  - El éxito del modelo en el negocio depende de ello.
  - Puede requerir intervención humana o no (volvemos a la *"evangelización..."*)
  - A-B testing es una buena práctica siempre que sea posible.

Retos del Machine Learning en la empresa

(... seguimos con la gobernanza …)

- ***Monitorización del modelo*** *(mi modelo está en Producción y lo estoy usando bien! Pero llevo un tiempo usándolo… Sigue siendo igual de bueno que cuando lo empecé a usar??)*
  - Backtesting de la puntuación:
    - Monitorizar su distribución en el tiempo.
    - Monitorizar su predictividad en el tiempo.
  - Backtesting de las variables del modelo:
    - Monitorizar su distribución en el tiempo.
    - Monitorizar su tendencia frente al score vs su tendencia frente a la variable objetivo.
  - A-B testing periódicos.

- ***Coexistencia de modelos*** *(tengo 20, 30, 50, 100 modelos… cómo controlo todo esto? Afectan unos modelos a otros?)*
  - Automatización de los informes del punto anterior → alertas automáticas cuando se detecten comportamientos anómalos.
  - Informes adicionales:
    - Matriz de correlación entre modelos (evol en el tiempo) → linealidad
    - Modelos como variables de otros modelos → no linealidad
  - A-B testing periódicos.
  - Versionado de modelos.

Retos del Machine Learning en la empresa

(... acabamos con la gobernanza … *o ella con nosotros?!* ;))

La mayoría de proyectos de ML que han fracasado o que han tenido bugs importantes y evitables han sido debidos a una gobernanza deficiente.

El hecho de no prestar suficiente atención a la gobernanza es debido al enfoque que se suele aplicar a ML: prueba-error, POC.

Este punto es de especial relevancia en sectores donde el uso de modelos está parcial o totalmente regulado, como la banca (riesgo de crédito) o los seguros.

La función de auditoría interna y/o de validación interna en toda empresa debería apoyar en este sentido al equipo de Analytics / ML / DS / … (pero normalmente no se meten).

Puede haber razones de momentum (inmadurez del ciclo, falta de desarrollo).

Incluso las herramientas comerciales existentes no han puesto todavía foco en esto (principalmente enfocadas en construcción, desarrollo, validación e implementación de modelos).

La conciencia del “Governance” en ML puede ser acelerada gracias a cambios que se avecinan a nivel legal, como la GDPR.

Retos del Machine Learning en la empresa

Datos: La gasolina de los modelos.

"Los datos están mal" *-- cómo de mal? -- todos o algunos? -- se pueden arreglar? -- aportan o destruyen valor?*

- ***Muchos datos***
  - Integrar datos de distintas fuentes: internas vs externas
  - Integrar datos en distintos formatos: tabular, texto plano, voz, imágenes, PDFs, formularios, web-scraping, …
  - Acceder a ellos de manera rápida (eficaz)
  - Siempre que merezca la pena usarlos (eficiente)
  - Cultura POC!
  - Enfoque en Admiral Seguros: la nada → bacos de variables → Data Lake?

- ***Buenos datos***
  - Imprescindible conocer la calidad de los datos.
  - Recurrir a ayuda… A veces los expertos en los datos no somos los DStists (no somos Dios) sino *"el negocio"*.
  - El poder de los datos… Y sus limitaciones!! *"Esto se empezó a preguntar hace 3 meses"*

- ***Tratamiento de datos***
  - Nuestro buen amigo Tidyverse ^_^
  - Librerías y/o frameworks de análisis descriptivos y visualización: plotly, ggplot2, … .
  - Keep it simple: a veces un simple str(), summary(), o table() son muy útiles.
  - Paquete reciente parecido a hmisc + histograma...

Retos del Machine Learning en la empresa

Modelización:

- ***Infraestructura y software***
    - Hasta dónde (y hasta cuándo) sirve el sw gratuito?
        - Volumen de datos.
        - Tiempos de ejecución.
        - Tamaño de equipo.
        - Restricciones de infraestructura:
            - Potencia.
            - Administración.
    - Opciones de pago
        - La referencia son los grandes líderes: AWS, MS (Azure), Google.
        - Otros clásicos: SAS e IBM.
        - Muchas más opciones disponibles (y cada vez más).
        - Todas con PROs y CONs.
        - Costes en función de uso, generalmente pequeños (y decreciendo).
        - Cloud vs On Premise: apuesta por cloud aunque probablemente dependerá del caso.
        - Valorar positivamente:
            - Elasticidad.
            - Escalabilidad.
            - Solución para todas las fases del ciclo de vida, o sólo para algunas??
            - Posibilidad de usar “raw software” desde estas plataformas.
            - Support.
            - Comunidad de usuarios.

Retos del Machine Learning en la empresa

Modelización:

- ***Aspectos técnicos***
    - Selección de atributos (variables relevantes).
        - El paso más relevante. El **principio de parsimonia** realmente tiene mucho valor!
        - Enfoque sencillo: usar cosas que ya existen!
            - Fscaret: wrapper de técnicas de selección de variables
            - Algoritmos que producen jerarquías de variables ellos mismos, como RF, XGBoost, ...
        - Enfoque complejo: algoritmo propio.
            - Debate: **de qué creemos que depende la importancia de una variable?**
    - Tuning de parámetros
        - Los que controlan el sobreajuste, o la complejidad, la tasa de aprendizaje, …
        - Randomized Search vs Grid Search vs Bayesian Optimization.
        - Pueden influir mucho en el poder predictivo.
        - Pueden provocar sobreajuste, cuidado!
        - En mi experiencia el paso previo tiene más impacto *(y yo antes pensaba lo contrario!!).*
        - Grandes aliados: caret, XGBoost, rBayesianOptimization, aunque cada vez hay más y mejores (h2o, …)
    - Técnica a usar
        - Expertise del DStist.
        - Tipo de problema.
        - Restricciones: implementación, negocio! *(recall: evangelizar…)*

Retos del Machine Learning en la empresa

Validación y aprobación:

- ***Infraestructura y software:***
  - Mismas consideraciones que en “Modelización”.
  - La solución de modelización y de validación debería ser la misma.

- ***Aspectos técnicos:***
  - Introducidos en “Gobernanza”.
  - Grandes aliados: los mismos que en la fase de “Datos”: Tidyverse, plotly, ggplot2, knitr y el mundo Markdown, shiny, ...

- ***Aspectos funcionales:***
  - Saber explicar qué se ha hecho con palabras no técnicas
    - El negocio no sabe (y casi siempre *no quiere saber* ML).
    - Fracaso garantizado si no sabemos explicar qué hemos hecho con palabras simples.
    - Entrenamiento con niñ@s… (no es broma del todo).

  - Un contraste de hipótesis: verdad o mentira?
    - **H0:** *“Un GLM es explicable y un modelo basado en RF, boosting de árboles o redes neuronales no lo es, son cajas negras”.*
    - Podemos hacer más gris (o blanca) la caja negra?
      - Distribuciones marginales (1D, 2D, 3D!) de las variables del modelo.
      - Ranking de variables.
      - Métricas de error: simulaciones, remuestreo.
      - Validar, validar, validar + explicar, explicar, explicar.
      - Demostrar: AB test.

Retos del Machine Learning en la empresa

Implementación: *he creado un monstruo… cómo lo uso?*

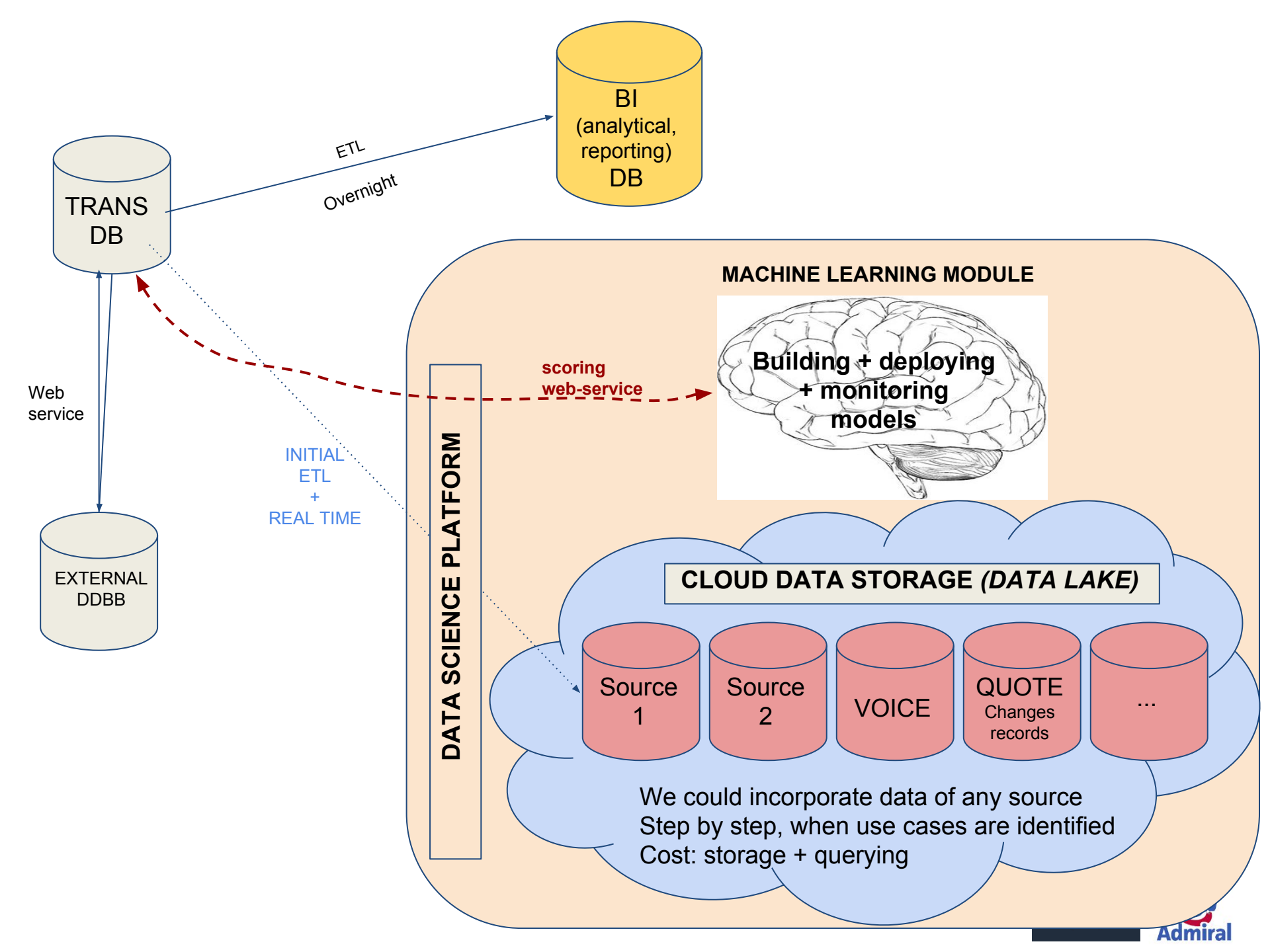
- ***Transcribir el algoritmo***
  - "Hago modelos en R pero en IT programan en Java".
  - Factible mientras hagamos modelos sencillos…
  - … O modelos complejos que usemos en modo "batch"

- ***Otras opciones:*** ver pág siguiente para un esquema tipo *(cocina propia, puede estar incompleto, el objetivo es sencillamente transmitir la idea)*.

  - Desarrollar una "calculadora de modelos"

  - Solución externa (plataforma de ML) que incluya un módulo de implementación ("deploy")

    En esencia son lo mismo, principales diferencias:
    - Soporte
    - Independencia
    - Coste
    - Escalabilidad


BI
(analytical,
reporting)
DB
ETL
Overnight
TRANS
DB
Web
service
EXTERNAL
DDBB
INITIAL
ETL
+
REAL TIME
MACHINE LEARNING MODULE
scoring
web-service
Building + deploying
+ monitoring
models
DATA SCIENCE PLATFORM
CLOUD DATA STORAGE (DATA LAKE)
Source
1
Source
2
VOICE
QUOTE
Changes
records
...
We could incorporate data of any source
Step by step, when use cases are identified
Cost: storage + querying

Retos del Machine Learning en la empresa

Monitorización:

Los principales conceptos los vimos en “gobernanza”.

Las plataformas de DS evolucionarán para incorporar módulos de este tipo, algunas ya los tienen.

La fase de validación (técnica) debe sentar las bases de esta fase: *la “primera iteración” de este bucle*.

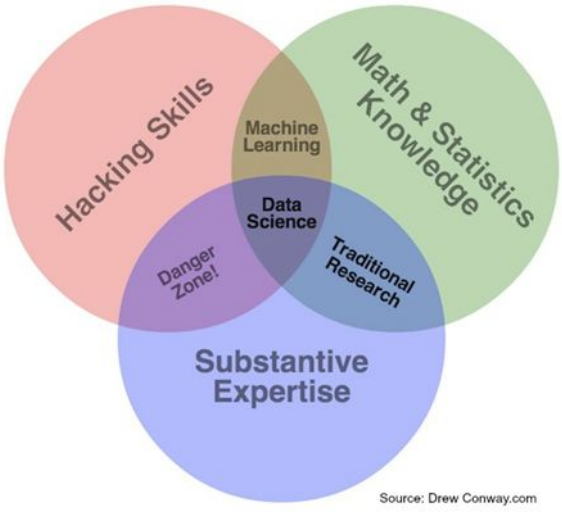
Automatización al máximo posible de informes *(el equipo de Analytics / ML / DS / … no puede tener infinitos recursos!)*

Comunicar rápido al negocio *(lenguaje simple!)* posibles anomalías, proponer acciones, anticiparse… sin “sobrerreaccionar”.

AB testing el mejor aliado para contrastar hipótesis que no se puedan dilucidar por otros métodos.

Retos del Machine Learning en la empresa

Equipo:



El activo más importante.

Recurso difícil de encontrar y caro.

Perfil multicomponente:

- Parte informática (hacker).
- Parte de negocio (experiencia, gestionar proyectos).
- Parte cuantitativa y analítica (científica-matemática-estadística).
- Trabajo en equipo:
    - Con el resto de miembros del equipo
    - Con el resto de personas de la empresa!

Enfoque “ideal”:

- Mínimo dimensionamiento,
- Máxima fidelización,
- Máxima automatización,
- Potenciar la “diferenciación”: *cuanto más distintos los perfiles, mejor* (aprenderemos los unos de los otros).

Stream Analytics

Cuándo puede ser conveniente?

- Contextos muy dinámicos, en los que de antemano sepamos que la realidad cambie con mucha frecuencia y por tanto los datos históricos pierdan valor cuanto más históricos sean.
- Cuando el poder predictivo del modelo no influye en la intensidad de una acción
  Ejemplo: modelo que reconozca caras a través de imágenes, o que transforme voz en texto, o un clasificador que “tokenice” textos. Una mejora fuerte de mi modelo por el hecho de que se reentrene con frecuencia es positiva siempre y contribuirá a mejorar todos los procesos que dependen del modelo.

Cuándo puede ser problemático?

- Cuando el poder predictivo del modelo influye en la intensidad de una acción
- Cuando necesite “datos limpios” con los que reentrenar el modelo
  Ejemplo: modelo que predice la probabilidad de fuga de clientes. Modelo actual:
  En el top 10% de población con puntuación más alta observo un 80% de fuga, y decidí dar a la gente que cae en este rango un descuento de hasta 50€ para evitar la fuga.
  Ahora mi modelo automatizado que se reentrena con nuevos datos él solo, en una de estas evoluciones me produce una ganancia fuerte de poder predictivo, y en el top 10% de población con puntuación más alta se observa un 95% de fuga → sigo queriendo dar un descuento de 50€? O doy más? O cambio el chip y los doy por “abandonados”?
  Por otra parte, si con el modelo actual estoy definiendo descuentos para todo el mundo, no tendré “datos limpios” para reentrenar un nuevo modelo: Dejar de dar descuentos durante un tiempo a todo el mundo, o bien definir un grupo de control al que nunca dé descuentos...

  En estos casos, veo preferible un sistema de reentrenamiento frecuente pero no automatizado, porque tiene que haber espacio para reflexionar sobre cómo cambiar las acciones de negocio derivadas del modelo.

Stream Analytics

Enfoques

- Desarrollo propio de un pipeline e infraestructura con un proceso tipo así *(ejemplo)*:
  - Tengo mi modelo entrenado ya, sea M_t0. Ahora, con una frecuencia f (horaria, diaria, semanal, …):
    i. Obtengo, limpio y preproceso datos y divido en train y test: Tr y Te
    ii. Entreno: M_t1
    iii. Mido error de test de M_t1 y M_t0 (ambos sobre la misma muestra, Te!), y:
      1. Si error(M_t1) < error(M_t0) actualizo el modelo
      2. En caso contrasio, no actualizo el modelo
    iv. Vuelvo al punto ii cuando llegue el siguiente instante f.

- Utilización de alguna herramienta de mercado que facilita este proceso a cambio de un coste. Una review de varias de las opciones existentes en este estudio de Forrester que se puede descargar desde aquí (hay que dar una serie de datos de contacto…).

¡ Gracias !

Néstor Alonso Cacheiro
nestor.alonso.cacheiro@gmail.com
nestor.alonso@admiral.es
@ACNestor
653743886